KLTE Számitástudományi Tanszék, Debrecen

Számitástechnika alkalmazása genetikai vizsgálatokban, különös tekintettel a szöveti antigénekre

Rochlitz Szilveszter

Bevezetés

A szöveti antigének kutatásával mintegy tiz éve foglalkozik az Országos Vérellátó Szolgálat Debreceni Alközpontjában egy kutatócsoport - Dr. Aszódi Lili és Dr. Stenszky Ernőné vezetésével. Az e téren elért eredményeik mind hazai, mind nemzetközi vonatkozásban ismertek. Ebbe a kutatómunkába kapcsolódott be a szerző hét évvel ezelőtt. E munkában csupán a legutóbb vizsgált problémákról és megoldásukról lesz szó, mellőzve az eredmények biológiai interpretációját.

A szöveti antigének alkotják a biológia egyik legösszetettebb, legpolimorfabb rendszerét. Vizsgálatuk jelentőségét kiemeli igen fontos szerepük a transzplantációs immunológiában. A legutóbbi vizsgálatok két fő irányba mutatnak:

a.) milyen mértékben mutat megegyezést az átlagos magyar populáció más populációval a szöveti antigének rendszerét illetően,

b.) bizonyos, eddig ismeretlen eredetü betegségek megjelenésének van-e kapcsolata a beteg szöveti antigén strukturájával.

Ezen kérdések eldöntéséhez feltétlen szükséges az egészséges magyar populáció és az adott betegségben szenvedők szöveti antigén-rendszerének megismerése és összehasonlitása.

1. Phenotipus megoszlás számitása

A szöveti antigének két diszjunkt halmazt alkotnak. Legyenek ezek

$$F = \{f_o, \ldots, f_n\} \; ; \quad f_i \neq f_{i'} \text{ ha } i \neq j$$

$$S = \{s_o, \ldots, s_m\} \; ; \quad s_g \neq s_h, \text{ ha } g \neq h$$

Egy ember 4 szöveti antigénnel rendelkezik:

$$\{f_i, f_j, s_g, s_h\} \; ; \quad f_i, f_j \in F \; , \; s_g, s_h \in S$$

s ezek két locusban helyezkednek el. Az első locuson az

$\{f_i, f_j\}$ a másodikon az $\{s_g, s_h\}$ antigén-pár.

Az f_o, ill. s_o az eddig még nem ismert antigéneket jelöli.

Definició: Az $\{a_k, a_l\}$ antigén-pár meghatároz egy phenotipust, ha $\{a_k, a_l\} \in F$

vagy $a_k, a_l \in S$ relációk egyike teljesül.

Definició: Egy antigén génfrekvenciáján a Hardy-Weinberg törvény alapján számitott

$$p = 1 - \sqrt{1 - r/N} \qquad /1.1/$$

értékeket értjük, ahol p az antigén előfordulásának gyakorisága egy N elemü mintában. A génfrekvencia ismeretében χ^2 -próba segitségével eldönthető, hogy két különböző populációban egy antigén előfordulási gyakorisága azonosnak tekinthető-e.

Az antigén struktura vizsgálatának genetikai szempontból is fontos kérdése, hogy egy adott mintában a különböző phenotipusok előfordulási gyakorisága megfelel-e a Hardy-Weinberg törvény értelmében várt értéknek?

Legyen $a_o,\ldots,a_n$ az egy locuson előfordulható antigének és $q_o,\ldots,q_n$ azok génfrekvenciái. Ekkor

$$\sum_{i=0}^{n} q_i = 1 \qquad /1.2/$$

egyenlőségnek kell teljesülnie, az a_o az eddig még fel nem derített antigéneket (blank), q_o ezek együttes frekvenciáját jelöli. Az öröklődés után a Hardy-Weinberg törvény értelmében

$$\left(\sum_{i=0}^{n} q_i\right)^2 = 1 \qquad /1.3/$$

Részletesen kiírva

$$q_o^2 + \sum_{i=1}^{n} q_i(q_i + 2q_o) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2\, q_i\, q_j \qquad /1.4/$$

Jelölje R_{ij} az $\{a_i, a_j\}$ phenotipus gyakoriságát. (Az a_i, a_j phenotipus szerológiailag nem különböztethető meg az a_i, a phenotipustól!) Az R_{ij} értékek ismeretében keressük azt a $Q'(q_1', \ldots, q_n')$ értéksorozatot, amely valamilyen értelemben a legjobban kielégíti az /1.4/ egyenletet. (Az /1.2/ alapján a q_o-t kifejezhetjük.)

A Q' becslést a maximum likelihood módszerrel határoztuk meg, miután a phenotipusok eloszlása polinomiálisnak tekinthető (1). Legyen

$$L(q_o, \ldots, q_n) = \overline{L}(Q) = \prod_{i=1}^{N} f(X_i Q) \qquad /1.5/$$

a likelihood függvény. Ekkor az

$$F_i = \frac{\partial \bar{L}(Q)}{\partial q_i} = 0 \qquad i = 1, \ldots, n \qquad /1.6/$$

egyenletrendszer megoldása adja a keresett becslést. Ezt az általánositott Newton-Raphson iterációs eljárással oldhatjuk meg.

$$Q'_{r+1} = Q'_r + I^{-1} F \qquad /1.7/$$

ahol Q'_i a Q' vektor az i-edik becslése, az F vektor komponensei az /1.6/-beli $F_1, \ldots, F_n$ értékek, az I^{-1} pedig az információs mátrix inverze, a variancia-kovariancia mátrix.

$$I_{ij} = - E\left(\frac{1}{\bar{L}(Q)} \frac{\partial \bar{L}(Q)}{\partial q_i} \frac{\partial \bar{L}(Q)}{\partial q_j}\right); \quad i,j=1,\ldots,n \qquad /1.8/$$

Miután polinomiális eloszlásról van szó,

$$I_{ij} = \sum_{\ell=0}^{K} \frac{1}{E(\ell)} \frac{\partial E(\ell)}{\partial q_i} \frac{\partial E(\ell)}{\partial q_j} \; ; \; i,j = 1,\ldots, n \qquad /1.9/$$

Az /1.8 és /1.9/-beli E függvényre

$$\sum_{\ell=0}^{K} E(\ell) = N$$

teljesül, ahol a szummáció az összes lehetséges phenotipusra vonatkozik.

Az /1.7/ iteráció pontossági feltételét χ^2 -próba segitségével adhatjuk meg:

$$\chi^2_n = (Q'_{r+1} - Q'_r)^T \; I \, (Q'_{r+1} - Q'_r) = F^T I^{-1} F \qquad /1.10/$$

Ha ez az érték kisebb egy előre megadott elfogadási szintnél, az iterációs eljárást befejezhetjük.

A számitások tényleges elvégzéséhez ki kell számitanunk az /1.5/, /1.6/ és /1.9/ formulákat.

$$L(Q) = \left(1 - \sum_{k=1}^{n} q_k\right)^{2R_{oo}} \prod_{j=1}^{n} \left(q_j \left(q_j + 2\left(1 - \sum_{k=1}^{n} q_k\right)\right)\right)^{R_{oj}} \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} (2q_iq_j)^{R_{ij}} \quad /1.5'/$$

$$F_i = \frac{-2R_{oo}}{1 - \sum_{k=1}^{n} q_k} + \frac{2R_{oi}\left(1 - \sum_{k=1}^{n} q_k\right)}{q_i\left(q_i + 2\left(1 - \sum_{k=1}^{n} q_k\right)\right)} -$$

$$- 2 \sum_{j=1}^{n} \left(\frac{-R_{ij}}{q_i} + \frac{R_{oj}}{q_i + 2\left(1 - \sum_{k=1}^{n} q_k\right)} \right); \quad i = 1,\ldots,n \qquad /1.6'/$$

$$I_{ji} = I_{ij} = \sum_{\ell=0}^{\frac{n(n+1)}{2}} \frac{1}{E(\ell)} \frac{\partial E(\ell)}{\partial q_i} \frac{\partial E(\ell)}{\partial q_j}; \quad i,j=1,\ldots,n \qquad /1.9'/$$

$$E(\ell) = \begin{cases} N\left(1 + \sum_{k=1}^{n} q_k\right)^2 & \text{ha } \ell = 0 \\ N\left(q_\ell + 2\left(1 - \sum_{k=1}^{n} q_k\right)\right) q_\ell & \text{ha } \ell = 1,\ldots,n \\ 2N\, q_k\, q_p & \text{ha } \ell = \frac{(2n-k-1)k}{2} + p, \quad \begin{matrix} k = 1,\ldots,n-1 \\ p = k+1,\ldots,n \end{matrix} \end{cases}$$

$$\frac{\partial E(\ell)}{\partial q_i} = \begin{cases} -2Nq_\ell & \text{ha } \ell = 1,\dots,i-1,i+1,\dots,n \\ 2N(1-\sum_{n=1}^{n} q_k) & \text{ha } \ell = i \\ 2Nq_p & \text{ha } \ell = \frac{(2n-i-1)i}{2} + p \text{ és } p=i+1,\dots,n \\ 2Nq_k & \text{ha } \ell = \frac{(2n-k-1)k}{2} + i \text{ és } k=1,\dots,n-1 \\ -2N\,(1-\sum_{n=1}^{n} q_k) & \text{ha } \ell = 0 \\ 0 & \text{különben} \end{cases}$$

Az /1.7/ iterációhoz szükséges egy induló értéket megadni, amit /1.1/ alapján célszerű számitani.

A számitásokat ALGOL programnyelven irt programmal a KLTE Odra-1204 számitógépén végeztük, több különböző minta esetén. A számolási idő átlagosan: n = 8 esetén (első locus) 1-2 perc, n = 14 esetén (második locus) 4-5 perc volt.

A keresett Q' értéksorozaton kivül egyszerű uton adódott még, hogy mely R_{ij} érték tér el jelentősen a számitottól, valamint a phenotipusok megbszlására vett minta statisztikai értelemben ellent mond-e a Hardy-Weinberg törvénynek.

2. Haplotipusok vizsgálata

Definició: Egy $\{a_i, a_j\}$ antigénpárt haplotipusnak nevezünk, ha

$$a_i \in F \quad \text{és} \quad a_j \in S .$$

A haplotipusok genetikai szerepe igen jelentős, mert egy családon belül az antigénpár nem válhat szét, igy egy ember csupán két haplotipussal rendelkezik, továbbá az utód mind az anyával, mind az apával egy-egy haplotipusban megegyezik.

A haplotipusok némelyike - a vizsgálati eredményeink alapján - összefüggést mutat a vizsgálatok tárgyát képező betegségek megjelenésével.

Definició: Haplotipus frekvenciának nevezzük az

$$x_{ij} = p_i P_j + D_{ij} \qquad /2.1/$$

értéket, ahol p_i az $a_i \in F$, P_j az $a_j \in S$ antigén génfrekvenciája, a D_{ij} pedig egy asszociációs együttható, a gametikus asszociáció mértéke.

$$D_{ij} = \sqrt{\frac{d}{n}} - \sqrt{\frac{(b+d)\,(c+d)}{n^2}} \qquad /2.2/$$

ahol a, b, c, d a két antigén együttes előfordulásának vizsgálatára készitett 2x2 kontingencia táblázat elemei, és $n = a + b + c + d$ teljesül.

Egyszerüsitő feltételek mellett meghatározhatók a varianciák is.

$$V(x_{ij}) \approx p_i^2 V(P_j) + P_j^2 V(p_i) + V(D_{ij}) \qquad /2.3/$$

és

$$V(D_{ij}) \approx \frac{1}{4N^3} r_i R_j$$

r_i az $a_i \in F$, az R_j pedig az $a_j \in S$ antigén előfordulási gyakorisága, a $V(p_i)$ és $V(P_j)$ értékek a mindkét locusra meghatározott Q' becsléssel együtt adódnak.

Az összes lehetséges haplotipus frekvencia kiszámitása mellett két kérdés foglalkoztatott bennünket:

a.) mikor fejez ki a D_{ij} érték csupán véletlenszerü kapcsolatot,

b.) a normál és beteg populációból vett minták összehasonlításakor a haplotipust alkotó két antigén társulási készségei között van-e lényegi eltérés.

Ezekre a kérdésekre legegyszerűbb megoldásként az χ^2 illetve a Fisher féle Z-próba segítségével kerestük a választ.

A számitások elvégzésére ALGOL nyelven készült program, amely a KLTE Odra-1204 számitógépén futott.

A haplotipusok vizsgálata még korántsem lezárt téma. Vizsgálataink jelenleg családok szintjén folytatódik, igy a ténylegesen előforduló haplotipusokat tudjuk meghatározni, s ez által közelebb tudunk jutni az eredeti biológiai problémák megoldásához.

Irodalom

(1) Bailey, N.T.J.: Introduction to the Mathematical Theory of Genetic Linkage, Oxford University Press, 1961.

(2) Yule, G.U.-Kendall, M.G.: Bevezetés a statisztika elméletébe, Közgazdasági és Jogi Kiadó, Budapest, 1964.

(3) Mattiuz, P.L.-Ihde, D.Piazza, A.Ceppelini, R.Bodmer, W.F.: New Approaches to the Population Genetic and Segregation Analysis of the HL-A System, Histocompability testing, Munksgaard, 1970.

(4) Sváb J.: A populációs genetika alapjai, Mezőgazdasági Kiadó, Budapest, 1971.

MTA SZTAKI

Azonositási kódok statisztikai vizsgálata

Garádi János

A kórházakban ápolt személyek azonositására általában a születési év, hó, nap, a nem, valamint az anya nevének kezdőbetüje szolgál.

Ezek az adatok jól használhatók, mivel nem változnak meg az ember élete során. A vizsgálatok azonban azt mutatták, hogy ezek az adatok önmagukban a személyek csak mintegy 30 %-át azonositják egyértelmüen. Ezért a fenti azonositó kódhoz hozzávehetjük az állandó lakóhely megyéjét és a település jellegét (pl.: falu, város, illetve Budapesten a kerület). Ezek az adatok egy éven belül a lakosság mintegy 2 %-ánál változnak meg. Ezen felül az olyan kódokat is hozzávehetjük, amelyeknek a változása a lakosságnak legfeljebb 3 %-át érinti.

A statisztikai vizsgálathoz modellként az ugynevezett cellabetöltési problémát használjuk.

Ebben a dolgozatban egy cellabetöltési problémával foglalkozunk, amely az összetartozó rekordoknak (emberek) véletlenszerüen kisorsolt azonositó számok segitségével történő azonositása során merült fel.

A mi esetünk annyiban különbözik a klasszikus cellabetöltési problémától, hogy a különböző cellákba esés valószinüsége különböző.

A feladat leirása a következő:

Adott N számu ember, és n számu különböző azonositószám,